Report

Part 1.regression

1. Feature preprocessing

I replace some value of budget by the average value of this column because there are some numbers which is extremely low, which is the dirty number.

And I use ast.literal_eval function to get the name of the first actor and director changing the data into json format.

2. Feature selection and evaluation

At first I choose some features related to the revenue based on the reality situation and put these features into the model, which is the top actor/actress, director , spoken language, runtime and the budget. And the MSR is 8.47E+15 , and correlation is 0.283447

And then I choose cast ,crew, budget, genres, language, company, countries these features, the MSR result is 1.05E+16, but I got a higher correlation score, which is 0.33821. and these score are based on the same model decision tree regression.

3. Model selection

The first model I try is linear regression model, but the score is not that well compared to the decision tree regression model. And the score is below:

MSR	correlation
9.14E+15	0.174886

Then I try the svm, And the score is below:

MSR	correlation
9.50386E+15	0.14

Then I try the decision tree, And the score is below:

MSR	correlation
1.05E+16	0.33821

And these score are based on the same selected features.

Part 2 classification

1. Feature preprocessing and Feature selection

The features which I selected and data preprocessing are same to the selection in part one.

2. Model selection

The first model I try is svm classification model, And the score is below:

average_precision	average_recall	accuracy
0.6925	0.5	0.6925

Then I try the dt classification model,which we can see it is clearly high performance than the svm models except the accuracy score, And the score is below:

average_precision	average_recall	accuracy
0.732443	0.585982	0.605

Then I try the SGDClassifier which implements a plain stochastic gradient descent learning routine. And the score is below:

average_precision	average_recall	accuracy
0.743596	0.608098	0.6325

And these score are based on the same selected features.

Summary

We can see that the classification is average better that the regression models, and the Stochastic Gradient Descent model has the best performance among the classification models wo predict the rate of movie